DOCUMENT RESUME

ED 422 366 TM 028 926

AUTHOR Witta, E. Lea; Daniel, Larry G.
TITLE The Reliability and Validity of Test Scores: Are Editorial Policy Changes Reflected in Journal Articles?
PUB DATE 1998-04-00
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Editing; Editorials; *Educational Research; *Reliability; *Research Methodology; *Scholarly Journals; *Scores; Statistical Analysis; Validity; *Writing for Publication
IDENTIFIERS Educational and Psychological Measurement

ABSTRACT

In 1994, the journal "Educational and Psychological Measurement" (EPM) instituted an editorial policy requiring authors to use technically appropriate language and methodological practices in their discussions of validity and reliability. To determine if this policy has had any effect on current publications, 150 validity and reliability studies were selected from 3 social science measurement journals ("EPM," "Psychological Assessment," and "Journal of Psycheducational Assessment") over a 3-year period (1995 through 1997). Language usage and methodological problems in these studies were compared to the same problems as reported in a previous study of 150 articles selected from the 3 volume years of the same journals immediately preceding the EPM editorial policy. Results indicate a statistically significant decrease in incidence of errors across time. Before the 1997 editorial, 108 of the 150 studies (72%) contained 1 or more errors, but after publication of the editorial, approximately 51% (77 of 150) analyzed contained 1 or more errors. (Contains 3 tables, 1 figure, and 10 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running Head: EDITORIAL POLICY CHANGES

ED 422 366

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
E. Lea Witta
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Reliability and Validity of Test Scores:

Are Editorial Policy Changes Reflected in Journal Articles?

E. Lea Witta and Larry G. Daniel

University of Southern Mississippi

Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 13-17, 1998.

TM028926

Abstract

In 1994, *Educational and Psychological Measurement* instituted an editorial policy requiring authors to use technically appropriate language and methodological practices in their discussion of validity and reliability. To determine if this policy has had any effect on current publications, 150 validity and reliability studies were selected from 3 social science measurement journals (*Educational and Psychological Measurement, Psychological Assessment,* and *Journal of Psycheducational Assessment*) over a 3-year period (1995 through 1997). Language usage and methodological problems in these studies were compared to the same problems as reported in a previous study of 150 articles selected from the 3 volume years of the same journals immediately preceding the *EPM* editorial policy. Results indicated a statistically significant decrease in incidence of errors across time.

The Reliability and Validity of Test Scores:

Are Editorial Policy Changes Reflected in Journal Articles?

Measurement soundness is essential to the integrity of social science research. Any study, however well designed, is suspect if information about the validity and reliability of the study's data is inadequate or absent (Qualls & Moss, 1996; Whittington, 1998). Researchers and professors of educational research often emphasize the importance of these characteristics, making the point that conclusions about validity and reliability of data should routinely precede substantive hypotheses based on the same data. Many times, however, these same scholars slip into the careless practice of referring to these important test *score* characteristics as characteristics of *tests*.

Reliability and validity are correctly conceived of as characteristics of test *data*, not characteristics of tests themselves. Nevertheless, the regular reader of social science literature will frequently see erroneous statements such as "the test is reliable" or "the validity of the test." Many researchers have recognized and attempted to correct this problem. For example, Wainer and Braun (1988) noted, "The 'validity of a test' is a misnomer" (p. 87). Popham (1995) added further, "Tests themselves, do not possess validity" (p. 40). This issue has also been addressed in the *Standards for Educational and Psychological Testing* (AERA,

APA, & NCME, 1985, p. 9): "Validity. . .refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores."

A related problem with language usage has to do with the tendency of some researchers to place more emphasis on the results of their validity and reliability studies than is warranted. For example, some authors erroneously claim that their results "prove" or "demonstrate" that a test is reliable. As is well known, statistical results do not "prove" anything; rather reliability or validity coefficients are estimates of score characteristics given the data in hand. If these estimates fall within reasonable ranges, researchers can have confidence that the scores on the instruments serve as relatively trustworthy measures of variables of interest.

In addition to language usage problems, there are at least two methodological problems that are used with some frequency in the reporting of reliability and validity results. First, some researchers have an affinity for reporting statistical significance tests along with their validity or reliability coefficients. In the case of a reliability coefficient, these statistical significance tests evaluate the null hypothesis that a set of scores is totally unreliable, a hypothesis that is meaningless considering that large reliability or validity coefficients may often be statistically significant even when based on extremely small samples (Thompson, 1994) whereas minute reliability or validity

coefficients will eventually become statistically significant if the sample size is increased to a given level (Huck & Cormier, 1996). Further, considering that reliability and validity coefficients are sample specific, statistical significance tests do not offer any promise of the generalizability of these coefficients to other samples.

A second methodological problem that is a cause of concern is the reporting of negative reliability coefficients. Conventional mathematical formulae for estimating reliability coefficients can yield negative values (Krus & Helmstadter, 1993); however, such values are not logically possible: test scores cannot be less than 0% reliable as such values would imply. Thus, negative reliabilities, even if they occur mathematically, are anomalous values that should not appear in research studies. Further, data that yield negative reliabilities should prompt suspicion.

In 1994, a new editorial policy requiring authors submitting manuscripts to *Educational and Psychological Measurement* (*EPM*) to use technically correct language when referring to test score characteristics (i.e., referring to the reliability and validity of scores or data rather than to the reliability and validity of tests) was implemented (Thompson, 1994). In establishing this policy, the editor noted:

> The subjects themselves impact the reliability [and validity] of scores, and thus it becomes an oxymoron to speak of "the reliability [or validity] of the test" without considering to whom the test was administered or other facets of the measurement protocol. . . . Therefore, the same measure, when administered to more heterogeneous or to more homogeneous sets of subjects, will yield scores with differing reliability [and validity]. . . . Based on these considerations, use of wording such as "the reliability of the test" or "the validity of the test" will not be considered acceptable in the journal. (pp. 839-841--emphasis added)

Daniel and Witta (1997) examined the use (and misuse) of language regarding score characteristics in 150 articles appearing in three social science measurement journals--*Educational and Psychological Measurement, Psychological Assessment*, and *Journal of Psychoeducational Assessment*--during the volume years encompassing calendar years 1992, 1993, and 1994. More than half of the articles surveyed contained two or more instances of inappropriate language. Other problems related to language misuse and various methodological concerns were also noted. By 1994, *Educational and Psychological Measurement* had a somewhat improved record, with instance of language misuse showing a steady decline over the three year period. By contrast, language misuse was rather

constant during the three year period for the three volumes from which the *Psychological Assessment* and the *Journal of Psychoeducational Assessment* articles were drawn.

Purpose

The purpose of the present study was to determine the degree to which the 1994 editorial policy initiated by *EPM* (Thompson, 1994) has had an effect on the number of errors relative to the reporting of validity and reliability results in articles published in the three journals reviewed by the Daniel and Witta (1997) study. Hence, the present study is a replication of Daniel and Witta (1997), but uses the three volume years--1995, 1996, and 1997--immediately following the publication of the *EPM* guidelines editorial calling for improved reporting of validity and reliability results (Thompson, 1994).

Method

As previously noted, three journals (i.e., *Educational and Psychological Measurement, Psychological Assessment,* and *Journal of Psychoeducational Assessment*) that regularly publish validity and reliability studies were selected as the source for the articles reviewed. Daniel and Witta's (1997) earlier study had used articles from these same three journals that were (a) selected from the volume years coinciding with the calendar years 1992 through 1994 and (b) accepted for

publication prior to Thompson's (1994) editorial call. The present study examined articles from the volume years coinciding with the calendar years 1995 through 1997, after publication of the editorial call. Results of the present study's analyses were compared to results of the Daniel and Witta (1997) study.

In selecting the population of articles for sampling, the following procedures were used:

1. All articles appearing in the "Validity Studies" section of *Educational and Psychological Measurement* over the 3-year period were identified.
2. All articles appearing in the main and "Brief Studies" sections of *Psychological Assessment* and all articles appearing in the main section of *Journal of Psychoeducational Assessment* for the 3-year period were scanned to determine whether they were primarily reliability/validity studies. Articles meeting this criterion were identified.

One hundred fifty articles were sampled from the resulting population of articles. Selected articles were coded on each of the following criteria (see Appendix A):

1. Was erroneous language implying the validity or reliability of a test used in the title, in the abstract, or in the body of the study?
2. Were statistical significance tests reported along with validity or reliability coefficients?

3. Was erroneous language used suggesting that findings had "proven" or "demonstrated" the validity/reliability of data/tests?
4. What type(s) of reliability evidence was (were) provided?
5. What method(s) was (were) used to assess validity?

Information from the coded data sheets was entered into SPSS/PC+ for analysis. The present study's data file was merged with the data file produced by Daniel and Witta (1997), resulting in an omnibus data file consisting of 300 coded articles ranging from 1992 to 1997, including 150 records (articles) accepted for publication prior to the Thompson (1994) editorial and 150 records (articles) published following the editorial. All articles were categorized by (a) the journal in which they appeared, (b) the time frame in which they were published (either prior to or after the editorial calling for change), and (c) the presence of language/methodological errors (having no errors vs. having one or more errors). The data were analyzed using a chi-square (χ^2) test of independence with correction for continuity to determine if there was a relationship between time of publication and incidence of errors.

Results and Discussion

There was a statistically significant relationship ($\chi^2 = 12.691$, df = 1, $p < .001$) between time and presence of error. Prior to the editorial (Daniel & Witta, 1997),

72% (108) of the 150 studies contained one or more errors. By contrast, approximately 51% (77) of the 150 studies analyzed after publication of the editorial contained one or more errors (see Table 1 and Figure 1).

Insert Table 1 About Here

Insert Figure 1 About Here

When data from each journal were analyzed separately, only the *EPM* data yielded a statistically significant relationship ($\chi^2 = 11.93$, df = 1, $p < .001$) between time and error. No statistically significant relationship between time and error was found for either the *Psychological Assessment* ($\chi^2 = 1.11$, df = 1, $p = .29$), or the *Journal of Psychoeducational Assessment* ($\chi^2 = .381$, df = 1, $p = .537$) data. Since more articles were selected from *EPM* ($n = 75$ per time frame as compared to no more than 50 per time frame for the other two journals), and since χ^2 is severely influenced by sample size, an additional test was performed to determine whether the initial statistically significant result for the *EPM* data might be a reflection of sample size: 50 of the *EPM* articles were randomly selected from the 75 articles in

each time frame and the data were reanalyzed. This analysis again resulted in a statistically significant relationship ($\chi^2 = 13.219$, df = 1, $p < .001$). Because only 25 articles were selected from the journal with the smallest number of studies for each time period, the sample size was again reduced for the *EPM* articles: only 25 articles were randomly selected from those from each time frame. A statistically significant relationship ($\chi^2 = 4.083$, df = 1, $p < .05$) between time and error was again detected. In each instance, the number of *EPM* articles containing one or more errors had decreased following the editorial.

Frequencies were also determined for each error category across time (see Table 2). It is refreshing to note that a negative reliability was not reported in any of the articles and that the erroneous use of language inferring that the results had "proven" the reliability or validity of the data or the test occurred in only three of the 300 articles reviewed. Furthermore, in most instances, the use of other inappropriate terminology had also decreased over time. There was a dramatic decrease, for instance, in the use of "the test is reliable" and "the test is valid" (or other equivalently erroneous language), with instances of language of these two types occurring 65 and 76 times, respectively, before the *EPM* editorial as compared to only 40 and 55 times, respectively, afterwards. Over 25% of the articles reviewed, however, still contained these statements.

Insert Table 2 About Here

Of the 300 studies reviewed, 47% investigated construct validity. Some of these used exploratory factor analysis (46%), some used confirmatory factor analysis (29%), and some used both procedures. Over half (50.7%) of the studies reviewed used internal consistency procedures as a method for obtaining reliability estimates (see Table 3).

Insert Table 3 About Here

Conclusions/Recommendations

The purpose of the present study was to determine if an editorial calling for improvement in the reporting of validity and reliability results had had an effect on the number of articles containing these errors in three measurement journals. Results indicated a statistically significant relationship between the number of articles with one or more errors and time of publication. All three journals had a reduction in the number of articles with one or more errors after the editorial was published. Unfortunately, however, only one journal, *Educational*

and Psychological Measurement, had a significant reduction in the number of articles containing errors, and that journal was the one in which the editorial appeared. Nevertheless, this result is encouraging in that it indicates that the editorial policy change at *EPM* has indeed been related to an actual change in editorial practice.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1995). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Daniel, L.G., & Witta, E. L. (1997, March). *Implications for teaching graduate students correct terminology for discussing validity and reliability based on a content analysis of three social science measurement journals.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 408 853)

Huck, S. W., & Cormier, W. H. (1996). *Reading statistics and research* (2nd ed.). New York: HarperCollins.

Krus, D. J., & Helmstadter, G. C. (1993). The problem of negative reliabilities. *Educational and Psychological Measurement, 53*, 643-650.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know*. Boston: Allyn and Bacon.

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement, 56*, 209-214.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21-37.

Table 1

Number of Errors by Journal Prior to and after the 1994 Call for Change

Number of Errors	Journal							
	Psychological Assessment (n = 50)		EPM (n = 75)		Psychoeducational Assessment (n = 25)		Total (N = 150)	
	Before	After	Before	After	Before	After	Before	After
1	10	7	17	15	6	7	33	29
2	7	10	9	5	6	5	22	20
3	6	3	10	8	5	3	21	14
>4	13	10	17	3	2	1	32	14
Total	36	30	53	31	19	16	108	77

Note. All sample sizes are given per time.

Table 2

Frequency of Inappropriate Terminology/Methodology by Category by Time

	Prior to		After	
Category	Count	%	Count	%
Title				
Test reliable/valid	42	28%	24	16%
Abstract				
Test is reliable	40	27%	22	15%
Test is valid	51	34%	38	25%
Study				
Negative reliability	0	0%	0	0%
Use of "Prove"	1	0%	2	1%
Test is reliable	65	43%	40	27%
p value reliability	6	4%	7	5%
Test is valid	76	51%	55	37%
p value validity	38	25%	11	7%

Note. Sample size = 150 for each group.

Table 3

Frequency of Use of Procedures for Deriving Reliability and Validity Estimates

Method	Count	%
Validity		
Content	20	3.3
Predictive	70	26.1
Concurrent	80	26.7
Construct	141	47
Exploratory Factor Analysis	137	45.7
Confirmatory Factor Analysis	87	29
Multi-trait/Multi-method	8	2.7
Convergent/Discriminant	80	26.7
IRT	7	2.3
Cross-Validation	5	1.7
Reliability		
Test-retest	50	16.7
Equivalent Forms	4	1.3
Split-Half	9	3
Internal Consistency	152	50.7
Inter-rater	25	6.3
Intra-rater	1	.3

Proportion With 1 or More Errors

by Journal and Time